1995

Test of ESL academic reading ability: the process of development and validity inquiry in an operational setting

Daniel James Harness
Iowa State University

Test of ESL academic reading ability:

The process of development and validity inquiry in an operational setting

by

Daniel James Harness

A Thesis Submitted to the

Graduate Faculty in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF ARTS

Department: English
Major: English (Teaching English as a Second Language/Linguistics)

Iowa State University
Ames, Iowa

1995

DEDICATION

To my wife:

Finally, an evening with no homework to do!

To my son:

Daddy's actually home tonight! Let's go play!

To my father:

Your love, patience and wisdom will never be forgotten.

To my mother:

Who has supported everything I have done—no matter how time-consuming.

To my siblings:

Just in case you thought I was in school only to put off getting a real job.

TABLE OF CONTENTS

INTRODUCTION

In the Fall of 1993, I examined the reading test that was being used by the English department for their English Placement Test (EPT) for incoming international students at Iowa State University (ISU). The EPT consisted of several subtests designed to measure listening comprehension, reading comprehension, vocabulary, and writing. The test was designed to determine which students needed further instruction in the previously mentioned areas. I examined the EPT for an assignment for the Principles of ESL Testing class I was enrolled in. Although no one was quite sure just exactly when the EPT was designed, the staff members that I talked to in the English department estimated the test had been in use since at least 1979. This meant the test had been in use for over fourteen years—it seemed to me as though it was time for a new test. There are two reasons for this line of thought. First, it is very possible that some where in the fourteen years of the test's use, some one has managed to acquire a copy of the test and has passed this along to other people. Second, the theory that this test is based upon is not known, so it is possible this test is based on an outdated theory of reading—or possibly on outdated theories of reliability and validity (reliability and validity will be discussed in more detail later).

Something else about the reading test caught my eye—the format. The following presents a typical question:

> Many university professors leave the classroom to become deans and department heads. Their new positions usually are more rewarding financially, but money is not their only reason for changing jobs. The modern educational system is organized in such a way that a major portion of academic recognition goes to the
>
> A. administrators.
> B. older teachers.

C. faculty members
D. trustees.

One problem with this format is the fact that the reading is too short. Rarely will students encounter such a short reading passage in their academic reading. It seemed to me that the passages used for testing reading should be longer. This would increase the authenticity of the readings (authenticity will be discussed in more detail later). It also seemed to me that the readings should contain reading material that international students are likely to encounter in texts and journals while studying at ISU, again to increase the authenticity of the passages. I decided at that time that I would take on the task of creating a new reading test for the English Placement Test. This new test would consist of passages from actual texts and journals used here at ISU. A variety of topics would be used to reduce the chance that a test taker could answer test items based simply on their knowledge of the topic and not on their reading ability. The passages would generally be around 400 words long and would be followed by five to six multiple choice questions about the passage. This is a format that is similar to the TOEFL, a test that most international students are familiar with. And like the TOEFL, there are time constraints on the testing situation which force limits on the length of reading passages and on the number of questions asked.

The original reading test included 35 questions, so I decided the new test would also contain about 35 questions. I wanted the test to measure how well students could locate the main idea of a passage as well as some of the supporting ideas. I also wanted the test to measure how well students could make inferences about the passage, i.e., I wanted to see how well students could use information in the passages to answer questions that were not directly related to the passages.

I presented this idea to a member of the faculty who further suggested that I discuss this project with the ESL Placement Coordinator. The ESL Placement Coordinator agreed that the old reading test was outdated and was in need of replacement and that my idea for a replacement was very sensible. At that time, we also decided to include vocabulary questions on the test. There was agreement among all of us that vocabulary was so important that it should be included on the test.

I borrowed some help from the ESL committee when putting the new test together. The ESL committee consists of several instructors who teach international students both in the English department and in the Intensive English Orientation Program. I passed out guidelines to the committee members for locating reading passages and for creating questions. After gathering passages and creating questions, the committee met to determine which passages and questions would be most appropriate for the new reading test. We created two versions of the test and gave the tests to international students (graduate and undergraduate) enrolled in the English 101C and 101D classes. After I had received the item analyses of the two tests, I met with the Placement Coordinator and a faculty member to decide which passages and questions we would use for the final version of the test that was to be used as part of a new EPT that was being designed.

The new EPT was given in January of 1995. Using the results of the reading test, the Placement Coordinator decided which students were in need of further reading instruction. It was left to me to examine the test results to inquire into the validity of the decisions that were made.

The purpose of the study at hand, then, becomes twofold. First, it is to examine the process used to develop the new reading test. Second, it is to examine

the process used to inquire into the validity of the decisions made using the results from the test. This will be done by a partial examination of validity which will focus on construct validity (this will not be a thorough examination of validity because a thorough examination is beyond the scope of the study at hand). Construct validity deals with how well a test measures what it is designed to measure. The discussion on validity will be followed by research on ESL reading (which for the most part branched off from research done on reading in English as a first language). After this will follow a more in-depth discussion of the process of the development of the test. This will be followed by a discussion of the results from the test and how these results can be used to provide construct validity evidence which is in turn needed to investigate the validity of test use. The conclusion of this study will include a discussion of what has been learned from this study and where we are to go from here.

LITERATURE REVIEW

J. B. Carroll, in 1968 (quoted in Bachman (1990)), defined a test in the following way: "a psychological or educational test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics" (p. 20), which Bachman restates as: "a test is a measurement instrument designed to elicit a specific sample of an individual's behavior" (p. 20). The behaviors being referred to are the responses that subjects give to test items. The characteristics being referred to are things such as intelligence, personality, language ability, and so on which will affect how subjects respond to test items. The WAIS-R and WISC-R, for example, are two popular tests that are intended as measures of specific characteristics—intelligence for adults and children respectively. How the subjects respond to the test items is their behavior. The Minnesota Multi-Phasic Personality Inventory is a diagnostic measure intended to allow users to infer possible personality disorders on the basis of test behavior. The TOEFL is intended as a measure of another characteristic—an individual's ability to use the English language. And these are just a few of the widely used tests on the market today.

The purposes of tests are as varied as the tests that have been developed. There is, however, a common rationale underneath every test's purpose—there is some behavior that it is intended to elicit. Although specific tests themselves elicit a wide variety of behaviors, the purpose of testing is less diversified. It is to elicit a behavior, often with the intent of inferring from the behavior a characteristic of an individual from which one might make predictions or comparisons. The variety of behaviors and the ways they are to be elicited and measured is what gives rise to the

variety of tests that have been developed. And what is to be elicited and how it is to be measured determines the design of a test.

Norm-Referenced vs. Criterion Referenced tests

Generally, tests such as the one that has been designed for the study at hand are designed for one of two purposes—to compare a subject's performance against his or her peers, or to compare a subject's performance against a set standard. An example of a test that compares a student against that student's peers is the TOEFL mentioned above. An example of a test that compares a student's performance against a set standard is a classroom final exam. Tests that are designed to measure a student's performance among a group of individuals are referred to as Norm-Referenced (NR) tests because the individual's test results are "interpreted with reference to the performance of a given group, the norm. The 'norm group' is typically a large group of individuals who are similar to the individuals for whom the test is designed" (Bachman, 1990, p. 72). NR test results provide a mean, or average, score with which an individual's test score is compared. Also, the individual's score can be used to compare that individual's performance with others.

On the other hand, Criterion-Referenced (CR) tests are designed to measure an individual's performance against a "criterion level of ability" (p. 74), in other words, a pre-set standard. Occasionally, a CR test is designed with a cut-off score which determines failure or success of a level of ability. The two most important distinctions between NR and CR tests are

> (1) in their design, construction and development; and (2) in the scales they yield and the interpretation of the scales. NR tests are designed and developed to maximize distinctions between individual test takers.... CR tests are designed to be representative of specified levels of ability...and the [test] items will be selected according to how adequately they represent these ability levels.... NR test scores are interpreted with reference to the performance of other individuals

> on the test, [but] CR test scores are interpreted as indicators of a level of ability or degree of mastery". (p. 75)

However, this doesn't mean that NR and CR tests are necessarily mutually exclusive. Tests can exhibit characteristics of both. The test designed for the study at hand, for example, is intended to exhibit characteristics of both NR and CR tests.

A brief introduction to construct validity

Problems do arise with testing. The most important question to arise is does a test measure what it claims to measure? This question is about construct validity. Tests are designed to elicit certain behaviors. The behavior that is of most interest in this study is test performance in a test of ESL academic reading. The question remains as to how accurately, if at all, the test that has been designed measures what it is designed to measure. One must consider the fact that human behavior may result from the interaction of multiple abilities. Therefore, it is extremely difficult to determine if the ability that has been measured with a test is actually the ability that it was intended to measure. There is no guaranteed method that will ensure that the ability that is desired to be measured is actually measured. Instead, the designer of a test must use a theoretical definition of the ability along with the empirical results of a test in order to determine whether the test has measured the desired ability. However, even with a high level of empirical support that the test measured the defined ability, there still is no guarantee.

The problem of construct validity that has been discussed here is the result of the unavoidable problem of errors being introduced into a testing situation. Errors can result from many different sources, from not properly or adequately defining the ability being measured to test questions that don't properly or adequately sample the ability being measured to distractions in the testing environment that

prevent subjects from giving their full attention to the test. Not adequately defining the ability being measured can introduce error by allowing the possibility of other (possibly unrelated abilities) to be included in the ability in question. Questions that don't properly sample the ability are likely to be sampling something else, which in turn leads to the introduction of error. Distractions can cause error due the fact that a test taker my have difficulty keeping her attention on the task at hand—the test could end up becoming a test of patience. Because of the questionability of the construct of a test (whether or not it is measuring what it was designed to measure) and the results it produces (are the results an indication of what was to be measured, or was there some other factor that produced the results), the presence of error is unavoidable. The goal then is to reduce the amount of error that is introduced into a test.

Authenticity and tests

The need to overcome the errors that are introduced into a test leads into another area of concern with tests, especially in language tests such as the one that has been created for this study. This is the issue of authenticity. Authenticity in a test is how well a test samples an ability as it would be used in a natural setting. This is important because it is desired for a test to measure as closely as possible the same ability that a student would use in a natural setting. Bachman (1990) discusses two approaches to defining authenticity:

> [T]he 'real-life' (RL) approach to defining authenticity essentially considers the extent to which test performance replicates some specified non-test language performance. This approach thus seeks to develop tests that mirror the 'reality' of non-test language use, and its prime concerns are: (1) the appearance or perception of the test and how this may affect test performance and test use ...and (2) the accuracy with which test performance predicts future non-test performance (predictive utility) .... The other approach...[is] the 'interactional/ability' (IA) approach [which] focuses on what it sees as the

> *distinguishing characteristic* of communicative language use—the interaction between the language user, the context, and the discourse. (p. 301-2)

Here Bachman is noting the importance of developing tests that replicate as closely as possible natural language and how it is used in a natural setting. It will be extremely difficult to replicate a 'real life' reading environment, after all, people choose to read in many different environments. We can, however, replicate the 'interactional/ability' by including on the test reading passages that replicate 'real life' reading material. The 'interactional/ability' that is being assumed for the design of the test for the study at hand is one in which students read passages of extended length from various sources, e.g., text books and journals. By designing a test that includes passages of extended length from texts and journals that are used at ISU, it is hoped that a setting similar to the natural setting described above has been achieved. This may reduce some of the error introduced by the testing process.

The need for reliability and validity

The problem of error in test scores discussed above is one reason why test scores must be examined. This is done by estimating the reliability of the test scores. Reliability in its simplest form is a "quality of test scores" (Bachman, 1990, p. 24). Reliability deals "with the consistency of...[a measure]...across different times, test forms, raters, and other characteristics of the measurement context". Since there is likely to be several sources of internal and external errors (such as those discussed earlier) in any measurement, it is strongly advised that test users determine sources of error and the effects these errors have on test scores. This is carried out by "making judgments based on an adequate theory of sources of error" (p. 24). Just how these sources of error affect test scores "is a matter of empirical research" (p.24).

Of further importance in this discussion is the issue of *internal consistency*. Internal consistency is concerned with how consistent test takers performances on different parts of a test are with each other (Bachman, 1990). It is a measure of how well test items are measuring the same construct. Inconsistencies in performance on different parts can be a result of several factors. For a reading test such as the one that has been designed for this study, possible sources of performance inconsistency may be caused by differing passage lengths or difficulty, inconsistent question formats, diverse passage topics, organizational pattern of passages or questions, etc.. Sources of error such as these may affect test performance randomly which may in turn affect test scores. So the more sources of error that can be identified and eliminated, the more consistent test performance should be and the more reliable test results will be.

Validity in its simplest form is the extent to which decisions made using test scores are meaningful, appropriate and useful (American Psychological Association, 1985). In general, a test that has been shown to be valid is valid only for the purpose for which it was designed to be used. If a test is used for a purpose other than its original purpose, the validity of the new test use must be evaluated. Because of the concern with measuring only what is intended to be measured, it is important when examining the meaningfulness of test scores to demonstrate that they are not affected by factors other than the ability being tested (Bachman, 1990) This sounds very much like reliability—which in fact it is. As such, it can be seen that reliability and validity are interrelated.

A high level of validity for a test used for its intended purpose is an indication that the possibility exists that the ability that was desired to be measured was actually measured and that unwanted factors have been reduced or eliminated.

Again, as was noted earlier, because we cannot be absolutely sure that what was measured was actually what was desired to be measured, we can only make inferences about results test results. Bachman (1990) further notes that if a test score is affected by errors, then it will not be meaningful and cannot be used a basis for valid interpretation or use. Subsequently, a test score that is not reliable due to the effect of errors cannot be valid. In other words, *reliability is a necessary condition for validity.*

The new view of validity

Recently, there has been a shift in researchers' view of validity. They are moving away from the old static view of a validated test—a test that has been proven valid for all time and can be used without further examination—to a new dynamic view of validity. This new view considers validation to be an ongoing process of gathering evidence and making justifications for test interpretations and uses. This process requires researchers and test users to examine closely the results of the measurement devices they use (Chapelle, 1994; Messick, 1989).

Central to the definition of validity is construct validity. Construct validity is the extent to which test performance is consistent with predictions that are made on the basis of a theory of abilities, or constructs (Bachman, 1990). In other words, it is the extent to which a test measures what is designed to measure. Messick (1980) describes construct validation as "a process of marshaling evidence to support the inference that an observed [performance] has a particular meaning" (p. 1015).

Using Messick's 1989 definition of validity, Chapelle (1994) discusses six different types of construct validity evidence: (1) *content evidence,* which is the "judgments of experts concerning the ability that test items measure"; (2) *empirical*

item analysis, which is the "observation and analysis of learners' performance on [a] test" as well as examination of "statistical item difficulty ([along with] other item statistics) and qualitative response analysis"; (3) *empirical task analysis,* which is the "qualitative research which probes the problem-solving processes of learners"; (4) *internal test structure,* which is "the extent to which patterns of empirical internal consistency realize theoretical expectations"; (5) *correlational studies,* which is the investigation of the relationships between performance on more than one test in addition to the "relationships between a test and behavior in other contexts"; and (6) *experimental research,* which allows examination of "hypotheses about test performance by systematically modifying test conditions to verify that observed test performance behaves in concert with theory-based predictions" (pp. 168-76).

For this study, content evidence, empirical item analysis, internal test structure, and experimental research will be examined. Content evidence will be examined by considering how well items on the test created for this study measure what they are supposed to measure. The empirical item analysis evidence will examine how well test takers performances on test items match difficulty predictions. Internal consistency, which is dependent on the internal structure of the test, will be used to provide evidence that test items adhere to the same structural relations as those hypothesized by the construct they represent. Experimental research evidence will be examined by comparing the performance of non-native speakers versus native speakers.

The new validity and the study at hand

The new view of validity will be used for the study at hand. This study will use the results of the reading test that has been developed to demonstrate the

content validity of the test. However, to argue the overall validity of the test, other pieces of evidence will need to be gathered. Messick (1989) argues that "[v]alidity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and action based on test scores" (p. 13). He notes that "validity is an evolving property and [that] validation is a continuing process" (p. 13). The rationales, according to Messick, include things such as relevance and utility of a test within its context, and value implications and social consequences of test use. The relevance and utility of a test refer to its usefulness. Value implications are associated with professionals opinions of the construct a test measures. The more agreement between test construct and professional opinion, the stronger the argument for validity. Social consequences are the effects that test use has on the context in which it is used. Graphically, the facets of validity can be portrayed as in Figure 1.

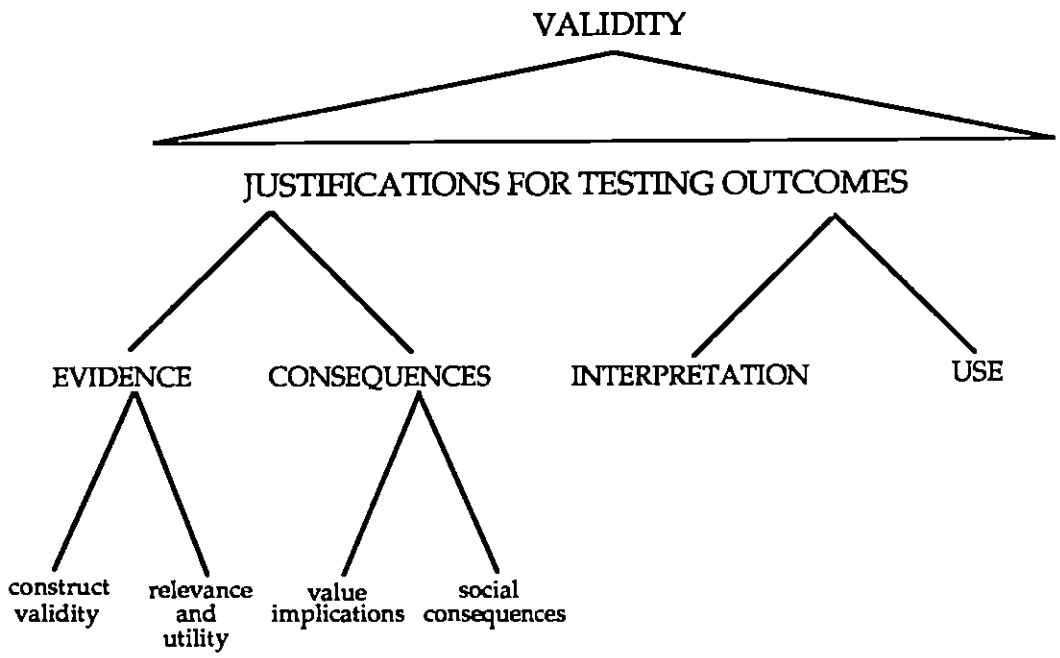

Figure 1. Messick's facets of validity (1989) as portrayed by Chapelle (1994).

The purpose of the test

The test I have developed for this study is a test of English as a Second Language (ESL) academic reading ability. The test is part of a three part battery that makes up the English Placement Test (EPT) at Iowa State University. The other tests included in this battery are a test of writing ability and a test of listening comprehension. This test is administered to all incoming graduate and undergraduate international students.

The purpose of the reading test is to give indications of incoming international students' ESL academic reading ability. The indications are used in order to single out those students whose reading ability may cause them trouble in their academic programs. These students are placed in a class designed to improve academic reading ability.

Towards defining the construct of the test

In order to investigate the construct validity of the test that has been designed, the construct the test is intended to measure must first be defined. For the development of the test, I theorized ESL academic reading comprehension as consisting of four separate components: vocabulary ability, the ability to identify main ideas, the ability to identify supporting ideas, and the ability to make inferences. These four components contributed to the construction of the test by determining what types of items would be used to test these characteristics. Four types of questions, each type designed to measure the different theorized components of the model, appear on the test. The overall test, then, actually consists of four subtests designed to measure each of the components of the model.

This model of ESL academic reading was theorized on research I have done in addition to what I teach my ESL students. Now I will examine the trends in reading comprehension both for English as a first language and English as a second language to see where the model I have created fits in with the research in reading comprehension.

I found that a review of the literature finds a seemingly infinite number of reading comprehension models both in English as a first language and English as a second language. Rather than review all of the models that have been created, I have chosen to review some of the approaches taken. This provides a more meaningful interpretation for the model of reading comprehension that I chose to develop for this test.

Reading comprehension (English as a first language)

A separation in the views on the concept of reading comprehension began in the early part of this century. The division was between those who considered reading to be a holistic general-factor process (e.g., Thorndike, 1917a, 1917b, 1917c) and those who considered reading to be a multiple-factor process (e.g., Gray, 1917) and this split exists up to this day. Holistic general-factor is a somewhat misleading name in that some of the theories that fall into this category actually hypothesize two components. Rost (1993) claims that there are researchers who take a middle-of-the-road approach dividing reading into two factors the first of which is associated with vocabulary knowledge and the second of which is associated with general reading comprehension or inferential reading. I, however, find little difference between the middle-of-the-road theories and the holistic general-factor theories, so I will consider these two approaches to be in one category which despite its

misleading nature I will continue to call holistic general-factor theories. What follows is a brief discussion of the two views of reading comprehension that exist today: the holistic general-factor process theory and the multiple-factor process theory.

The best place to start the discussion of the views of reading comprehension is with the work of Thorndike (1917a, 1917b, 1917c). He analyzed errors produced by students who were asked to answer questions about short paragraphs they had just read. He saw the evidence for two levels of processing. The first was a word level process and the second was a process he referred to as reasonings. He held the opinion of many others at that time that reading was a "compounding of habits" (p. 323), but he refused to state that reading was the combination of subskills. Instead, his evidence suggested to him that more than one level of processing was involved with reading.

Thurstone (1946) in his factor analysis of a nine component model of reading comprehension proposed by Davis (1944) found that a single factor (reading ability) could account for variance in the nine subtests that Davis used in his test of reading comprehension. Davis argued that they tested nine different skills; however, Thurstone claimed his study indicated a lack of evidence supporting Davis' model. R. L. Thorndike (1973-74) used factor analyses, correlational studies, and stability studies of difficulty in reading test items under translation from one language to another to demonstrate that a wide variety of tests seemed to show only two processes involved with reading—reasoning and decoding. These results were in support of the findings by E.L. Thorndike in 1917. Vacca (1980) did a study of approaches to teaching reading. His study indicated that students who received instruction in the subskills approach did no better than students who were taught

the holistic approach to reading. He concluded that for pedagogical purposes there appears to be no argument for splitting reading into subskills. Vernon (1962) in his study of 108 male subjects from the U.S. and Britain found through a factor analysis that two main processes could account for a majority of the variance observed—a vocabulary ability and a reading ability. Pettit and Cockriel (1974) did a study of sixth grade students who were given two tests of reading, one a test of literal reading consisting of six subskills and the other a test of inferential reading consisting of five subskills. A factor analysis revealed that the two tests were in fact measuring two distinct factors. Just as in the Vernon study, a pattern of two processes accounting for most of the variance emerged.

In contrast to the holistic theorists were theorists like William Gray (1917) who felt that reading was the combination of many different components. They felt that these components were separate from one another and could be tested individually. Davis (mentioned above) after a survey of literature hypothesized that reading involves nine separate subskills:

1. Knowledge of word meanings.
2. Ability to select the appropriate meaning for a word or phrase in the light of its particular contextual setting.
3. Ability to follow the organization of a passage and to identify antecedents and references in it.
4. Ability to select the main thought of a passage.
5. Ability to answer questions that are specifically answered in a passage.
6. Ability to answer questions that are answered in a passage but not in the words in which the question is asked.
7. Ability to draw inferences from a passage about its contents.
8. Ability to recognize the literary devices used in a passage and to determine its tone and mood.
9. Ability to determine a writer's purpose, intent, and point of view, i.e., to draw inferences about a writer. (p. 186)

Johnson (1949) in her survey of literature available at that time also hypothesized the existence of several different components emerging in reading comprehension.

Some of the more common components she noted were vocabulary ability, inferential ability, and organizational skills. Davis (1972) in a survey of reading tests found many of these tests tested several common skills. The components that were listed most often were vocabulary, inferential ability, understanding statements that support central thought (supporting ideas), and recognizing central thought (i.e., the thesis). Spearitt (1972) designed a test to investigate the eight subskills that Davis proposed in 1968. In this model Davis included recalling word meanings; drawing inferences about the meaning of a word from context; finding answers to questions stated explicitly or in paraphrase; weaving together ideas into the content; drawing inferences from the content; recognizing a writer's purpose, attitude, tone, and mood; identifying a writer's techniques; and following the structure of a passage. A factor analysis revealed that most of the variance could be accounted for by four factors: recalling word meanings; drawing inferences about the meaning of a word from context; recognizing a writer's purpose, attitude, tone, and mood; and following the structure of a passage.

A review of the multi-factor theories reveals that many common traits exist in these theories: vocabulary knowledge, inferential ability, organization and structural knowledge, understanding statements that support the main topic, and grasping the writer's purpose. Some of these are components that are also seen in the model of reading comprehension used for the study at hand.

Reading comprehension (English as a second language)

The history of reading study in English as a second language is not quite so lengthy as it is for reading in English as a first language. The importance of reading in English as a second language began to be seen in the late 1960s with the sudden

increase of ESL students, both here in the U.S. and in Britain. The research conducted drew heavily from research that was being done in reading English as a first language and this is still true today (Grabe, 1991). Researchers such as Goodman (1967) and Smith (1971) led the way. The models that these researchers were proposing were drastically different from what had been seen in the English as a first language research of the preceding seven decades. These models were based on a new psycholinguistic approach to reading. According to Grabe (1991), Goodman proposed that reading is more than a "process of picking up information from the page letter-by-letter" (376-77). It is a "selective process" (377) meaning that readers don't examine words in detail; instead, they bring background knowledge to the reading and predict information, sample the text and confirm predictions. This approach to reading is what is known as the *top-down* approach because it focuses on higher levels of mental processing. Researchers noted that the pure form of top-down would require no text, so these also became known as interactive approaches because they actually hypothesized an interaction between information in the text and the background knowledge of the reader. In opposition to these were the earlier models that had been used—models that focused on letter and word recognition. These models are known as *bottom-up* models.

Researchers such as Clarke and Silberstein (1977) and Coady (1979) saw the importance of including the lower level mental processes when instructing ESL reading students. Approaches that combine bottom-up and top-down processing are also called interactive by Grabe (1989, 1991). The models of reading that have resulted from the numerous theories is almost countless and would fill an entire book (for a review of a few of the major models, see Samuels & Kamil (1984)).

As was mentioned earlier, much of the research that is being done today in ESL reading comes out of research done in reading English as a first language. Several things have become clear from both English as a second language and English as a first language reading. For one thing, the models created are numerous, and for another, the process of reading is complex and difficult to understand. This has lead many researchers back to trying to identify components of the reading process (as was done in research earlier in the century). Grabe (1991) mentions that research has lead to the proposal of "six general component skills and knowledge areas:

1. Automatic recognition skills
2. Vocabulary and structural knowledge
3. Formal discourse structure knowledge
4. Content/world background knowledge
5. Synthesis and evaluation skills and strategies
6. Metacognitive knowledge and skills monitoring" (p. 379)

Automatic recognition skills are skills used in identifying shapes of letters and words. This skill is referred to as automatic because the reader is unaware of the use of these skills. Vocabulary and structural knowledge is the knowledge of the meanings of words and knowledge of how words can be put together to form discourse. Formal discourse structure (or formal schemata) knowledge is knowledge of how a whole text is organized, for example expository, comparison-contrast, etc. Content/world background knowledge (content schema) is the prior knowledge text-related information that a reader brings to a text. Synthesis and evaluation skills and strategies are those skills and strategies used to "evaluate text information and compare and synthesize it with other sources of

information/knowledge" (381). It is the synthesis and evaluation skills in particular that help a reader to make predictions about later text development and organization. Metacognitive knowledge and skills monitoring consists of "knowledge about cognition and the self-regulation of cognition" (382). According to Grabe, "Knowledge about cognition, including knowledge about language, involves recognizing patterns of structure and organization, and using appropriate strategies to achieve specific goals....As related to reading, this would include recognizing the more important information in a text; adjusting reading rate;...using search strategies for finding specific information;...and so on" (382).

A vague resemblance exists between the component skills seen here and the component skills seen in earlier models such as those proposed by Davis in 1944 and 1968. In the components that Grabe outlines, however, much more emphasis is placed on the reader and the knowledge a reader brings to a text. This, of course, is a result of the shift to the more interactive approach to reading comprehension.

Before concluding the discussion on reading comprehension, one last issue will be briefly discussed—the issue of the transfer of first language reading skills to a second language. Research has been done to examine the transfer of skills from a first language to a second language (see e.g. Koda (1988) and Singleton and Little (1991)), but no definite conclusions have been made as to what skills are transferred and to what extent these skills are transferred.

The model I have chosen

As was mentioned in the introduction of this paper, the model of ESL academic reading that I have chosen derives from personal teaching experience. In my ESL classes, I have found it necessary to teach determining the main idea,

locating supporting ideas, and making inferences about a reading. I have on occasion found teaching a little vocabulary to also be of help to my students. For this reason, I have formulated a four component model of reading comprehension that consists of vocabulary, identifying main ideas, identifying supporting ideas, and making inferences. Research exists to support the inclusion of all four of these components in a model of reading comprehension.

One area of agreement among all the theories, whether holistic or multi-factor, is that vocabulary is one of the processes involved with reading. In addition to the role of vocabulary in a theoretical definition of reading, there is an abundance of evidence to indicate that vocabulary is one of the components of reading comprehension. Vocabulary ability is knowing the definition or meaning of a word. Several studies have found a strong relationship between vocabulary and reading ability. Laufer (1992), for example, found a high correlation between lexical (vocabulary) level in a second language (L2) and L2 reading ability. Lewis (1987) found vocabulary to be a good predictor of reading performance. Also, a review of the research of reading comprehension models I have done reveals the consistent view that reading involves vocabulary knowledge.

Evidence also exists to support including identifying main ideas as a part of reading comprehension. The main idea is the thesis of a reading passage. The many books on teaching reading in English as a first language that I have reviewed (Alexander (1979); Alexander and Heathington (1988); Dechant (1991); Howards (1980); Johns (1986); and Kennedy (1977) to name just a few) demonstrate the need to include identifying main ideas as a component of reading comprehension.

There is also evidence to support including identifying supporting ideas as one of the components in reading comprehension. Identifying supporting ideas is

the ability to identify information that is important to the understanding of a reading passage. Evidence is seen in the previously mentioned books on teaching reading which all demonstrate the need for identifying supporting ideas as a component of reading comprehension. And more evidence for including identifying supporting ideas is seen in the review of multi-factor theories. Several of these theories include identifying supporting ideas as one of the main components of reading comprehension ability.

There is a great deal of evidence to support inferencing as one of the components of reading comprehension. Inferencing is the ability to use information provided in a reading passage to answer questions not directly related to the passage. Olsen in his 1985 study of third-graders found that good readers were better at answering inferential questions than were poor readers. Davey and Macready (1985) and Singer (1988) found much the same results—good readers show a better ability to answer inferential questions than do poor readers. A final argument for including inferencing as a component is the fact that the many of the models discussed so far include inferencing as a component of reading comprehension.

The research used to support the model I am using is research in reading in English as a first language. But, as was mentioned earlier, much of the research done in ESL reading is based on the research done in reading in English as a first language. Similarly, I am using research in reading in English as a first language to support my model for reading in English as a second language.

The textbooks used to teach reading that I examined (see above) all suggested first teaching vocabulary, then identifying main ideas, then identifying supporting ideas, and finally making inferences. These textbooks also suggested teaching

grammar. This was usually recommended after teaching vocabulary. However, since we do not have a grammar class in the English department (this is taught in the Intensive English Orientation Program), grammar was excluded from this test.

The order of teaching just mentioned seems to imply certain levels of difficulty for each of the components of the test, and students performances on the test should indicate this. Students should perform best on vocabulary items and worst on inferencing items. This prediction of how students should perform on items is referred to as an item difficulty prediction. Item difficulty predictions will be used later to investigate for the construct validity of the test.

Much of the research supporting my model that has been discussed so far is English as a first language research. Now I will examine how my model fits in with current research in reading in English as a second language. I will examine how my model fits in with both the interactive models (I am speaking of models that are both top-down and bottom-up) and the six components that Grabe speaks of (discussed earlier in this section).

The model I have chosen exhibits characteristics of both top-down and bottom-up processing. The top-down processing is seen in the inferencing component. Inferencing will be influenced by the knowledge that the subjects bring into the texts that they read. The bottom-up processing is seen in the vocabulary component. Recognizing a word and recalling the meaning of the word are bottom up processes. Identifying main ideas and supporting ideas fall in the middle of the two processes. Knowing how a text is organized does seem to have some effect on how well readers can recall text (see e.g. Conner (1984) and Carrell (1984)). Knowing how a text is organized may help readers to better recognize where the main idea and where supporting ideas will occur in the text. This can be seen as an interaction

of two processes: previous knowledge of possible textual organizations that the reader has brought to the test, and recognizing a particular organization. It would seem, then, that my model could be considered an interactive model—several components acting and interacting together.

Where do the components I have selected fit in with components currently considered to be important to the reading process? The Vocabulary component is an easy fit. It fits right in with the vocabulary component that researchers consider to be so important.

Where does the Main Idea component fit in? Where this and the Supporting Ideas component fit in isn't exactly clear. It is possible they might fit in with the formal discourse component since recognizing the organization of a text may aid readers in locating main and supporting ideas. They could to some extent be included in the metacognitive knowledge and skills monitoring component, too, since recognizing important information in a text is part of the knowledge of cognition.

Where does the Inferencing component fit in? It would probably best fit in with the synthesis and evaluation skills and strategies component. Synthesis and evaluation skills and strategies consists of the reader's ability to evaluate the text and make predictions about the information presented. These are the important parts in inferencing, which is the ability to use information provided in a reading passage to answer questions not directly related to the passage .

It would appear, then, that the model I have chosen fits into both the interactive and the component views of reading comprehension.

Evaluation of reading tests

Generally, the research concerning the evaluation of reading tests has not been favorable. For the most part, studies have done little to support the interpretation of reading test scores. Aronson and Farr (1988) note that it is highly important to remember that answering reading test questions is only an *indication* of reading comprehension. As a result, reading comprehension is not actually displayed—it is inferred from test results. Readence and Moore (1983) further note that in standardized reading tests, examinees aren't allowed to justify answers or pose questions of their own—going back to the question of authenticity and, therefore, test takers aren't in a natural setting. Sternberg (1991) strongly argues that reading tests are not good indicators of real world reading ability. He provides a table (see Figure 2) that identifies twelve differences between reading for a test and reading in school and everyday life. Though some of his ideas are excessive and/or questionable, several (for example, the first three and the very last) are valuable considerations in testing reading. By examining the differences from Sternberg's analysis, it becomes clear that reading tests are poor examples of authentic language use. Lyons (1984) continues the assault on reading tests from a different approach. He argues that the construct validity of many criterion-referenced standardized reading tests is low by suggesting that since the components that make up the reading process are not highly agreed upon, the construct of many reading tests (which are based on these components) is in question. He further argues that extreme caution should be used when interpreting results from standardized reading tests.

	Standard tests	School/everyday life
1.	Passages are short.	Passages are moderate to long.
2.	Learning from reading is massed.	Learning from reading is distributed.
3.	Recall is immediate.	Recall is delayed.
4.	Recall is entirely intentional.	Recall is largely incidental.
5.	Comprehension is based on a single type of question, usually multiple choice.	Comprehension is based on multiple types of assessments.
6.	The reasoning in the passage is very tight.	The reasoning in the passage is variable and often loose.
7.	Assessments measure evaluation of arguments.	Assessments measure construction as well as evaluation of arguments.
8.	Reading passages tend to be emotionally neutral.	Reading passages tend to be emotionally charged.
9.	Reading passages are often unmotivating and boring.	Reading passages are often motivating and interesting.
10.	Reading situations minimize distractions.	Reading situations contain many distractions.
11.	Evaluations are for a single purpose.	Evaluations are for multiple purposes.
12.	Students do the reading because they have to.	Students often (but not always) do the reading because they want to.

Figure 2. Sternberg's differences between reading for standardized tests and reading for school and everyday life (1991, p.541).

Freedle and Kosten (1993), however, leave aside that problems of authenticity and construct definition and argue construct validity from a different perspective. They were able to demonstrate through predictions of item difficulty how evidence for construct validity for some multiple choice reading tests (they used the TOEFL reading test as an example) can be gathered.

It is important to note that Freedle and Kosten used predictions of reading item difficulty as evidence to argue the construct validity of the TOEFL reading test.

Predictions of item difficulty is one of the types of construct validity evidence that was mentioned earlier (and is one of the types of construct validity evidence that will be used in this study). As can be seen, Freedle and Kosten were drawing upon the new view of validity that has emerged—the view that validity is argued by gathering evidence to justify testing outcomes.

The construct evidence that Freedle and Kosten examine is empirical item analysis done through an examination of item difficulty predictions. The study at hand will go somewhat beyond the Freedle and Kosten study and will examine three additional forms of construct validity evidence. As was mentioned earlier, this study will also look at content evidence, internal test structure, and experimental research. These will help to investigate the construct validity of the test. However, in order to demonstrate the validity of the test, other pieces of evidence will ultimately need to be investigated. As was mentioned earlier, these are relevance and utility, social consequences, and value implications. Although these pieces of evidence were not investigated for this study, they will be discussed in terms of how they could provide evidence for the validity of the test.

METHOD

Part of the purpose of this study is to discuss the development of a test of ESL academic reading ability based on a model of ESL academic reading comprehension which includes the following:

1. Ability to recognize vocabulary meanings in context.
2. Ability to identify the main idea.
3. Ability to identify supporting ideas.
4. Ability to make inferences from a text.

The context for which the test was developed required a means of identifying incoming international students who need further reading instruction from those who don't. The purpose of the test is to provide indications of students' ability so that decisions can be made concerning which students need further reading instruction.

This section will discuss the subjects used for the study and the procedure used to develop the materials. This section will also discuss the procedure used for giving the test and for analyzing the results.

Subjects

The subjects for the pilot tests were graduate and undergraduate international students enrolled in their first semester at ISU. The 92 students who volunteered were all enrolled in ESL academic writing courses. A total of six sections of 15-16 students took part; two sections were graduate students and the rest were undergraduates. The test was given during the last two weeks of class during the Fall 1994 semester.

The subjects for the final test were graduate and undergraduate students enrolled in their first semester at ISU. These 172 students all took the English Placement Test (EPT) of which this test was a part. The first test was given the week before the Spring 1995 semester. A make up version of the EPT which contained the same version of the reading test was given during the first week of the Spring semester.

The native English speakers who took the test consisted of 42 freshmen enrolled in first and second semester freshmen composition classes in the Spring of 1995. These subjects took the test during class time at the end of the semester.

Materials

The pilot tests. To create the pilot tests, a list of guidelines (see Appendix A) was passed out to the ESL committee which consists of several ESL instructors. The guidelines explained the process for gathering reading passages and for creating test items. The passages gathered were all readings from actual texts and journals used at ISU that members of the ESL committee felt were similar to those students would likely encounter in the near future. The purpose for using these materials was to create a representative sample of reading materials international students may encounter at ISU. By presenting authentic texts to the students, I intended to reduce some of the error introduced into the testing situation.

A total of twelve passages were gathered and from these a total of seven were chosen by the ESL committee for the pretests: (1) Peter Paul Rubens, Flemish Artist (from a journal called *Art and Civilization*) is a passage about a Flemish artist named Peter Paul Rubens; (2) Cyclamen in all Their Infinite Variety (from a journal called *The Garden*) is a passage about a variety of flower called the Cyclamen; (3) The Haber

Process (from a textbook called *Chemestry: The Central Science*) is a passage about a process for synthesizing ammonia directly from nitrogen and hydrogen; (4) The Effect of Laughter on the Human Body (from a textbook called *Walk, Amble, Stroll: Vocabulary Building Through Domains*) is a passage which discusses research done on the effects of laughter; (5) Plants as Therapy (from a book called *Plants as Therapy*) is a passage about horticulture therapy which is essentially using gardening as a treatment for certain emotional and physical problems; (6) Culture Shock (from a textbook called *Principles of Language Learning and Teaching*) is a passage discussing a problem suffered by many individuals living in a culture other than their own; and (7) Laboratory Research on Warnings (from a textbook called *Psychology*) is a passage that discusses research done on warnings. For each of these passages, questions were created to test the four identified components: vocabulary, identifying main ideas, identifying supporting ideas, and the ability to make inferences (see Appendix B for a sample passage with questions). The seven passages and their questions were split into two test versions each containing five passages. Three passages were repeated on both versions of the test and the other four were split between the two. Attention was given to creating two tests that contained as closely as possible the same difficulty level and amount of reading. Version 1 contained passages 1, 2, 5, 6, and 7 with a total of 36 questions. Version 2 contained passages 3, 4, 5, 6, and 7 with a total of 38 questions.

The Test Evaluation Center at ISU did item analyses on both versions of the test. The item analyses of the pre-tests are given in Tables 1 and 2.

Item difficulty is the number of individuals correctly answering an item divided by the total number of individuals answering the item. Item discrimination is a measure of how well an item separates the good test takers from the bad. Brown

(1990) suggests that acceptable item difficulties range between .3 and .7 (p. 119). He also suggests that item discriminations that fall between .3 and .39 are considered reasonably good with some room for improvement, and he considers .4 and above to be very good item discriminations (p. 120).

Both versions of the test produced results we were hoping for. Both tests had good internal reliability estimates and average scores for both tests were around 75%. Since the subjects taking the test had all scored 500 or better on the TOEFL, we expected average scores on this test to be fairly high.

The final test. After receiving the item analyses, I met with the ESL Placement Coordinator and one of the professors in the English department to determine which passages would be on the final version of the test. We wanted a test that would take approximately 40 minutes for test subjects to complete. We decided that the final version of the test should contain four passages as opposed to five passages as was used in each of the pilot tests. We chose passages that as a whole contained as many items as possible that met the guidelines set by Brown. We found that the last four passages of pilot test version 2 best met these requirements. We realized that a few of the items were outside of Brown's guidelines. Some had rather high item difficulties (well above .70 indicating the items may have been too easy) and others had item difficulties that were rather low. But overall, we felt the items were testing the abilities they were designed to test. Since we were under a time constraint, we chose to leave the items as they were. We decided to wait until results of the final test came in to re-examine the items under question. The final test included passages 4, 5, 6, and 7 and a total of thirty questions.

Table 1. Pilot test Version 1 item analysis

Number of items = 36	Mean score = 21.84	KR-20 = .78
Item	Item difficulty	Item discrimination
1	.58	.28
2	.91	.28
3	.96	-.09
4	.76	.25
5	.40	.11
6	.07	.06
7	.82	.15
8	.98	.37
9	.40	.15
10	.40	.20
11	.47	.09
12	.38	.20
13	.29	.36
14	.89	.50
15	.71	.45
16	.82	.58
17	.91	.29
18	.57	.40
19	.69	.39
20	.42	.19
21	.51	.15
22	.68	.43
23	.89	.42
24	.80	.47
25	.34	.34
26	.80	.19
27	.82	.57
28	.86	.49
29	.66	.36
30	.74	.86
31	.66	.75
32	.14	.36
33	.70	.82
34	.79	.90
35	.82	.93
36	.55	.71

Table 2. Pilot test Version 2 item analysis

Number of items = 38	Mean score = 26.02	KR-20 = .86
Item	Item difficulty	Item discrimination
1	.94	.42
2	.89	.45
3	.96	.11
4	.98	.35
5	.66	.11
6	.87	.20
7	.21	.15
8	.96	.00
9	.98	.34
10	.68	.19
11	.89	.42
12	.89	.42
13	.59	.15
14	.96	.26
15	.98	-.12
16	*	*
17	.87	.77
18	.91	.50
19	.89	.75
20	.54	.55
21	.60	.56
22	.56	.53
23	.49	.14
24	.56	.35
25	.73	.56
26	.91	.72
27	.47	.38
28	.79	.37
29	.74	.60
30	.79	.58
31	.79	.67
32	.78	.94
33	.72	.80
34	.38	.54
35	.68	.84
36	.74	.94
37	.65	.75
38	.51	.54

*Answer was keyed incorrectly

Procedure

As was mentioned previously, this test was given as part of the English Placement Test (EPT). The EPT was given in a large lecture hall during the mid-morning. The students had already completed a 30-minute writing sample and a 40-minute listening comprehension test before they took the reading test. Each student was given a test packet that included the reading test and a computer answer form. After receiving instructions, students read the passages and questions and responded by filling in the appropriate answer for each test item. The students were allowed 40 minutes to complete the test though most students finished in 30.

For the sake of comparison, the reading test was also given to native speakers. The native speakers took the test during class time and were allowed 35 minutes. Most completed the test in under 25 minutes. Test results and item analyses for non-native and native speakers appear in the Results and Discussion section of this paper.

Analysis

The analysis of the results included an examination of the internal consistency estimates (the KR-20s) which gave an indication of internal test structure, and item difficulty predictions (discussed above) which served as an empirical item analysis. Also included in the analysis was a judgmental examination of how well test items measured the constructs they were intended to measure (content evidence), as well as comparisons between native and non-native speakers which served as experimental research. These were used to investigate the construct validity of the test.

RESULTS AND DISCUSSION

The internal structure of the test will be examined by observing the internal consistency reliabilities for the overall test and for subtests composed of items representing each of the components of the model. Internal consistency reliabilities were determined by an item analysis. The empirical item analysis will be done by examining performance on each of the components and comparing to the performance that was theorized. This will be done by observing item difficulty scores for each of the components. The content evidence will be examined by discussing how well professionals involved with the test feel the test items are measuring what they are intended to measure. Experimental research will be done by comparing the performance of native speakers to that of the non-native speakers. This will represent the evidence needed to investigate the construct validity of the test. But first, the overall results from the test will be discussed.

The overall results (non-native speakers)

The distribution of the overall test scores (see Figure 3) indicates that the distribution is similar to an NR test thus justifying using NR item analysis procedures. The figure also demonstrates what is known as a negative skewing. This means that many of the scores are clustered to the right side of the distribution. The reason why this happened is due to the fact that all the subjects who took the test scored at least 500 on the TOEFL (this is the minimum score required to be admitted into ISU). Therefore, it is expected that many of the subjects will do well on this test. Distributions for each of the components are also shown (see Figures 4-7). The same type of distribution should be seen for each of the components as was seen for the overall test, and for the same reason. As can be seen in Figures 4, 6 & 7

this is what happened, at least for vocabulary, identifying supporting ideas, and inferencing. It is difficult to make any conclusions regarding the distribution for identifying main ideas (see figure 5) since there was only four items, although the distribution seen takes on a form similar to the distributions seen for the other components. Perhaps if there had been more main idea questions, the distribution would have taken the same shape as seen in the other components.

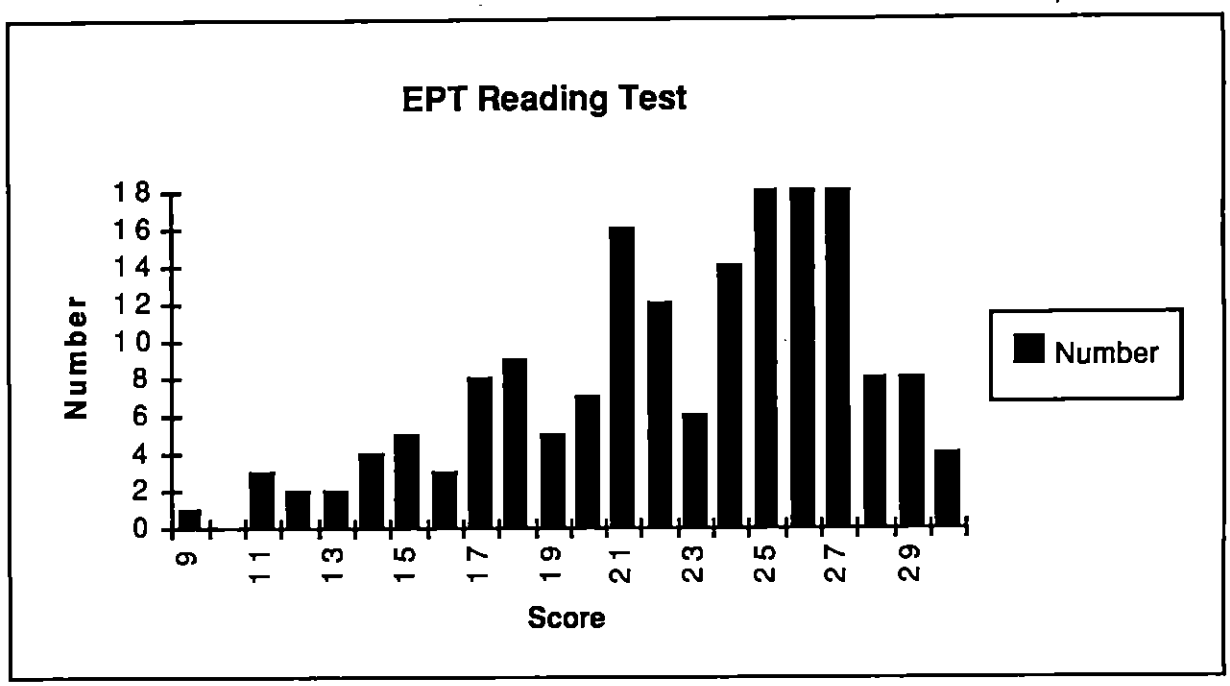

Figure 3 English Placement Test reading scores (non-native speakers)

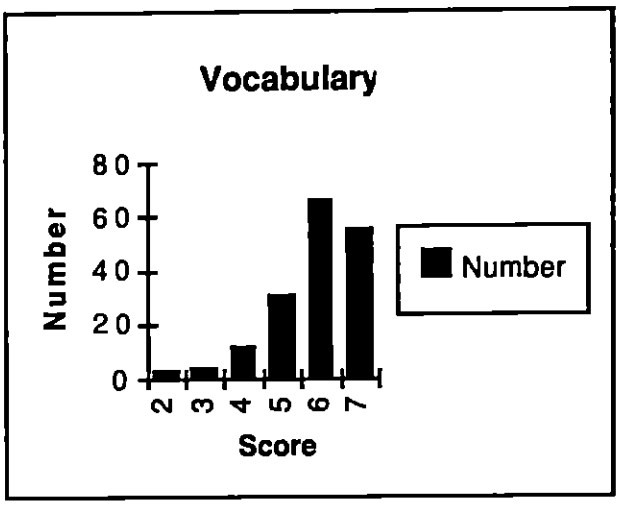

Figure 4 Vocabulary component scores (non-native speakers)

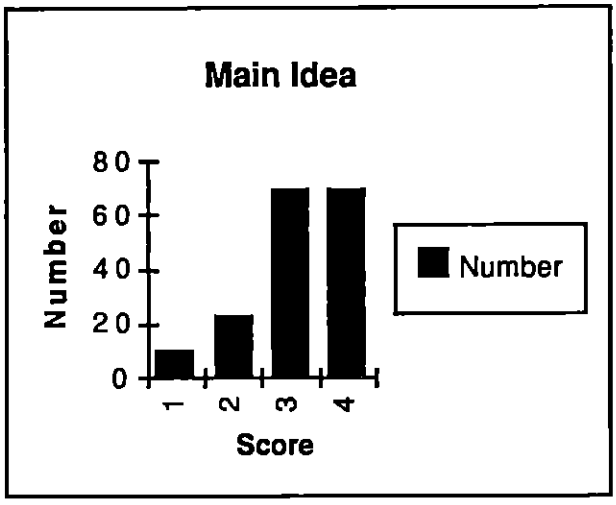

Figure 5 Main Idea component scores (non-native speakers)

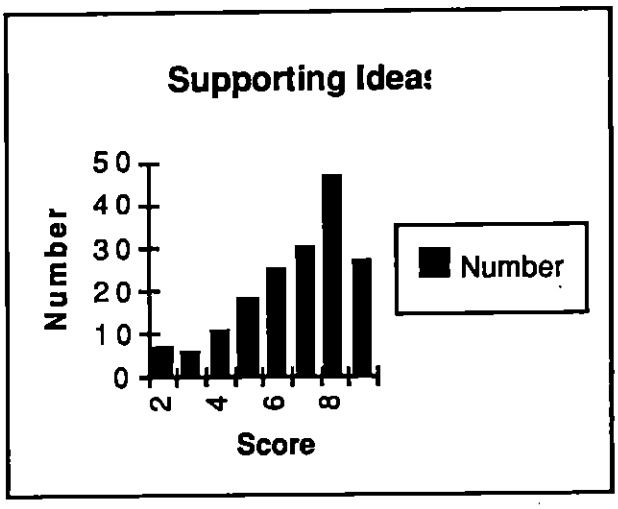

Figure 6 Supporting Ideas component scores (non-native speakers)

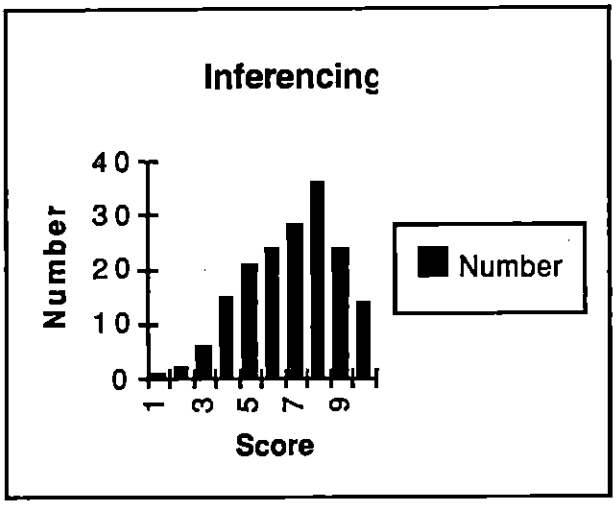

Figure 7 Inferencing component scores (non-native speakers)

The outcome from an item analysis of the test results for non-native speakers appears in Table 3. Item analyses were also done for each of the components of the model. These were done by clustering together items associated with each of the four components of the model and treating each of the clusters as an individual test.

Table 3. Item analysis results (non-native speakers)

Mean score = 22.56 (75%) KR-20 = .81		
Item	Item difficulty	Item discrimination
1	.98	.09
2	.71	.51
3	.90	.42
4	.93	.42
5	.54	.36
6	.96	.24
7	.98	.21
8	.79	.39
9	.82	.45
10	.87	.54
11	.92	.27
12	.71	.44
13	.74	.40
14	.54	.41
15	.46	.38
16	.51	.50
17	.86	.35
18	.87	.43
19	.45	.45
20	.72	.30
21	.73	.46
22	.81	.35
23	.82	.82
24	.85	.38
25	.68	.54
26	.39	.32
27	.82	.60
28	.89	.42
29	.71	.43
30	.65	.41

Results from the item analyses for each of the components appear in Tables 4-7. The reliability estimates for each of the components were adjusted for a 30 item tests using the Spearman-Brown Formula (see Thorndike and Hagen (1977)). Because of the way internal reliability estimates are figured, the more items that are on a test, the more accurate the internal reliability estimate will be. The number of test items on the subtests ranges from 4 to 10. To achieve a better approximation of internal reliabilities for each of the subtests, reliability estimates were adjusted for 30 item tests (the same number of items as on the overall test).

The overall results indicate a wide range of item difficulties, although many are in the range desired (between .30 and .70). And for the most part, item discriminations fall in the desired range (.30 and above). Although it seems evident that some items may be too easy, it does appear that the items are discriminating well between those who did well on the test and those who didn't.

What follows now is a discussion of the four pieces of construct validity evidence that will be used for investigating the construct validity of the test.

Content evidence

The first step in arguing for construct validity will be in examining content evidence. Content evidence is the judgments of experts concerning the ability that test items measure. Each of the test items is designed to measure one of four hypothesized components of reading ability—identifying vocabulary in context, identifying main ideas, identifying supporting ideas, and making inferences. It was experts (who were members of the ESL committee) who created the test questions and all were satisfied that the questions were testing what they were created to test. Also, the experts who created the final version of the test were in agreement that the

items included on the final version of the test were testing what they were designed to test. There was consensus, then, among those who created and implemented this test that the test items were testing the individual hypothesized components of ESL academic reading. It is important to note, though, that this consensus was among those who created and implemented the test. Experts outside of test creation and implementation would have to be questioned in order to complete the investigation of the content validity of this test.

Internal test structure

The next step in arguing for construct validity is an examination of the internal structure of the test. This was done by examining the reliabilities achieved with this test. The overall results from the non-native speakers indicate that the test has a fairly high level of internal consistency (KR-20 = .81). This indicates that the items seem to all be measuring the same construct. Although several of the items have rather high item difficulty level (a high item difficulty level means the item may have been too easy), most of the item discriminations fall within the guidelines suggested by Brown (mentioned earlier). The internal consistency estimates for each of the components are respectably high as well. The observed reliability for the vocabulary component was .43, for the main idea component it was .35, for the supporting idea component it was .65, and for the inferencing component it was .57. When adjusted for 30 item tests, these rose to .76, .80, .86 and .80 respectively (see Table 8). Again, these are all acceptable levels of internal consistency indicating that items in each of the subtests are reliably measuring the same construct as other items in the same subtest.

Further content evidence could be gathered by correlating each of the subtests with similar tests that are established measures of each of the hypothesized components of ESL academic reading. A high level of correlation between the subtests and other established measures would further strengthen the argument that each of the subtests is testing a different component of ESL academic reading ability.

Table 4. Item analysis of vocabulary items (non-native speakers)

Mean score = 5.86 (83%) KR 20 (adjusted for 30 items) = .76		
Item	Item difficulty	Item discrimination
4	.93	.42
6	.96	.24
7	.98	.21
11	.92	.27
14	.54	.41
20	.72	.30
27	.82	.60

Table 5. Item analysis of main ideas items (non-native speakers)

Mean score = 3.15 (79%) KR 20 (adjusted for 30 items) = .80		
Item	Item difficulty	Item discrimination
1	.98	.09
9	.82	.45
16	.51	..50
24	.85	.38

Table 6. Item analysis of supporting ideas items (non-native speakers)

Mean score = 6.70 (74%) KR 20 (adjusted for 30 items) = .86		
Item	Item difficulty	Item discrimination
2	.71	.51
3	.90	.42
10	.87	.54
12	.71	.44
17	.86	.35
18	.87	.43
21	.73	.46
25	.68	.54
26	.39	.32

Table 7. Item analysis of inferencing items (non-native speakers)

Mean score = 6.85 (68%) KR 20 (adjusted for 30 items) = .80		
Item	Item difficulty	Item discrimination
5	.54	.36
8	.79	.39
13	.74	.40
15	.46	.38
19	.45	.45
22	.81	.35
23	.82	.82
28	.89	.42
29	.71	.43
30	.65	.41

Empirical item analysis

The next step in arguing the for construct validity will be an empirical item analysis. The texts that I examined (see above) which are used to teach reading assume a specific order for learning each of the abilities measured in this test. Generally, the pattern is to teach vocabulary first followed later by teaching to recognize main and supporting ideas followed by teaching to make inferences from texts. There seems to be levels of difficulty associated with each of these components. The easiest level appears to be vocabulary—the most difficult is inferencing. It follows, then, that the subjects should do best on the vocabulary component and worst on the inferencing component. Although item difficulties varied greatly within each component (see Tables 4-7), the average scores for each of the four components (see Table 8) are consistent with predictions about how subjects would perform on each of the components. It is seen that the test takers had more success correctly answering vocabulary items than main idea items. They had more success correctly answering main idea items than they did supporting idea items. And they had more success correctly answering supporting idea items than they did inferencing items. These results provide further evidence for the construct validity of the test.

Table 8. Summary of components (non-native speakers)

Component	Mean	KR-20 (adjusted for 30 items)
Vocabulary	5.86 (83%)	.76
Main Idea	3.15 (79%)	.80
Supporting Ideas	6.70 (74%)	.86
Inferencing	6.85 (68%)	.80

Experimental research

The final piece of evidence used to demonstrate the construct validity of the test is experimental research. This was done by comparing the results of native speakers with non-native speakers.

The results for the native speakers was rather surprising. It was expected that native speakers would outperform non-native speakers. Instead, the native speakers scored somewhat lower; however, a t-test indicated that the difference was not statistically significant. An item analysis of the test results appears in Table 9. The distribution of scores (see Figure 8) and item difficulties and discriminations were very similar to non-native speakers. Interestingly though, the native speakers did not follow the same pattern concerning the components (see Tables 10-13 for a component by component breakdown of item difficulties and Table 14 for a summary of the components). For native speakers, the main ideas component saw the highest score (80%) followed by the vocabulary component (77%), the inferencing component (68%), and the supporting ideas component (67%). Part of the reason for this was the fact that both the vocabulary and the supporting ideas components had some items with unusually low scores which brought the averages for these two components down. When these items are excluded, much the same pattern occurs for both native and non-native speakers.

Are the results from the native speakers a cause for alarm? Actually, no. There are at least a couple of possible reasons for the results that were seen:

1. The native speakers were not highly motivated to do well on the test. Many native speakers hurried through the test simply to finish quickly in order to be able to leave class early. For example, one of the native speakers achieved a score of only four. This is a good indication that some of native speakers were applying little, if

any, effort. It is possible, therefore, that some native speakers simply guessed at many of the items instead of trying to figure out correct answers.

2. The non-native speakers consisted of both graduate and undergraduate students. And of the undergraduate students, several already had one, two, or more years of college experience. Also, international students at ISU tend to be the higher achieving students from their countries. The native speakers were all undergraduates in their first or second semester of college. They were a more eclectic mix consisting of both high and low achieving students. A better matched set of non-native and native speakers (for example, if the non-native speakers were all freshmen, too) may produce results more like what was originally expected—native speakers outperforming non-native speakers.

All of these factors added together indicate that although more research should be done, there is no cause for alarm for the results that were seen.

One interesting result does appear. The native speakers did worse on the vocabulary component than did the non-native speakers. Subsequently, they did worse on the overall test. This is what would be expected after examining the research I did earlier. Whether or not this was pure coincidence is not clear.

It would appear, then, that the experimental evidence doesn't support test validity—at least on the surface. Additional experimental research needs to be done. Further experimental research may provide better evidence for construct validity. For example, the test could be given first without the passages (passage-out), and then with the passages (passage-in). The differences in scores could be compared to determine how much of an effect including the passages has on test scores. Statistically significantly higher passage-in scores would provide strong experimental evidence that the test is actually testing reading ability.

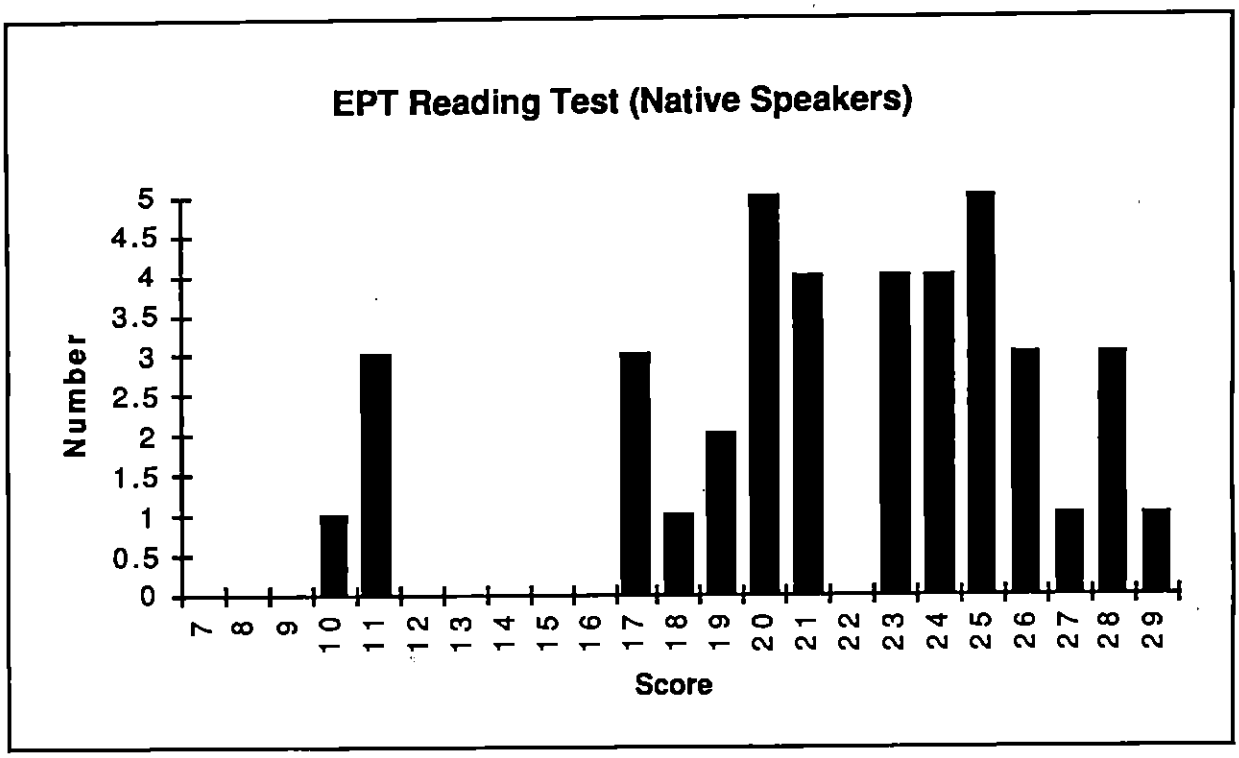

Figure 8 English Placement Test reading scores (native speakers)

Table 9 Item analysis results (native speakers)

Mean = 21.2 (71%) KR 20 = .86		
Item	Item difficulty	Item discrimination
1	.95	.59
2	.59	.24
3	.98	.03
4	.80	.66
5	.85	.59
6	.85	.27
7	.95	.57
8	.90	.74
9	.71	.57
10	.93	.47
11	.90	.33
12	.29	.33
13	.73	.28
14	.63	.28
15	.29	.40
16	.71	.52
17	.88	.35
18	.76	.38
19	.66	.37
20	.49	.46
21	.71	.30
22	.63	.46
23	.73	.58
24	.73	.40
25	.63	.61
26	.28	.29
27	.49	.46
28	.77	.65
29	.69	.74
30	.59	.51

Table 10. Results from vocabulary items (native speakers)

Item	Item difficulty	Item discrimination
4	.80	.66
6	.85	.27
7	.95	.57
11	.90	.33
14	.63	.28
20	.49	.46
24	.79	.80

Table 11. Results from main idea items (native speakers)

Item	Item difficulty	Item discrimination
1	.95	.59
9	.80	.66
16	.71	.52
24	.73	.40

Table 12. Results from supporting idea items (native speakers)

Item	Item difficulty	Item discrimination
2	.59	.24
3	.98	.03
10	.93	.47
12	.29	.33
17	.88	.35
18	.76	.38
21	.71	.30
25	.63	.61
26	.28	.29

Table 13. Results from inferencing items (native speakers)

Item	Item difficulty	Item discrimination
5	.85	.59
8	.90	.74
13	.73	.28
15	.29	.40
19	.66	.37
22	.63	.46
23	.73	.58
28	.77	.65
29	.69	.74
30	.59	.51

Table 14 Summary of components (native speakers)

Component	Mean score
Vocabulary	5.41 (77)%
Main Idea	3.19 (80)%
Supporting Ideas	6.05 (67)%
Inferencing	6.84 (68)%

The first three pieces of evidence gathered provide strong evidence for the construct validity of the test. The performance of the native speakers, at least on the surface, seems to contradict the construct validity of the test. However, it is hard to determine if the poor performance of the native speakers was due to the construct the test was based on or on factors other than the test takers' language ability. The native speakers were not well matched to the non-native speakers and they had nothing to lose by performing poorly on the test. A similar pool of native and non-native speakers and a similar level of motivation may likely see an increase in the scores for native speakers.

Limitations on the construct validity of the test

The construct validity that has been investigated is applicable only to this particular testing situation. When the test is given again, the process of gathering evidence to investigate construct validity will need to be done all over again. Several issues contribute to this line of reasoning. First, the subjects taking the test will be different, which will result in different internal consistencies which could effect the appearance of the internal structure of the test. A group that produces lower reliability scores (which are measures of internal consistency) could bring into question the internal structure of the test since the internal consistency of the test provides evidence for the internal structure. Second, the experts using the test may change their opinions of what ESL academic reading consists of; or for that matter, the experts using the test may not be the same experts. This would reduce the value of the decisions made. Third, further research may indicate that the performance of both the non-native and native speakers was due to something other than academic reading ability. Further research may indicate the test is actually measuring something other than academic reading ability, which could explain why native speakers did worse than non-native speakers. As can be seen, because of all of the changes that can occur from one testing situation to another, a test that has shown to be valid for one testing situation cannot be assumed to be valid for another.

CONCLUSION

The purpose of this paper has been twofold. First was an examination of the process of developing the test that was used for this study. Second was an examination of the process of gathering construct validity evidence. For this particular study, four pieces of evidence were examined: content evidence, empirical item analysis, internal test structure, and experimental research.

As was demonstrated in this paper, two important considerations must be observed when creating and validating a test. First is the context in which a test will be used and second is the purpose of the test. For the study at hand, the context of the test was one in which international students who need further instruction in ESL reading need to be separated from those who don't. The purpose of the test was to give an indication of students' ESL academic reading ability so that decisions could be made concerning which students are in need of further reading instruction.

Test development process

Once the context and purpose of the test were established, a hypothetical model as a representation of ESL academic reading ability. From this model, passages were selected and items were created that would test the individual components of the model.

Construct validity inquiry process

From test results, conclusions were made concerning the internal structure of the test. In addition to this, predictions concerning item difficulties were confirmed. These became two of the three pieces of evidence that were used to demonstrate the construct validity of the test. The third piece of evidence was an examination of

content evidence. Although the level of agreement between experts associated with the test development process was not investigated, there was some consensus among the experts that the test items seemed to be testing the constructs they were supposed to. The final piece of evidence was an examination of experimental research. The test was administered to native speakers with the expectation that the native speakers would perform better. This wasn't the case—non-native speakers out-performed native speakers. It was theorized that the results of the native speakers may not be a good indication of their actual academic reading ability. Although the first three pieces of construct evidence gathered justified the construct validity of the test for academic reading ability, the fourth piece of evidence not did not.

But, in order to demonstrate the validity of the test, other pieces of evidence have to be investigated. The relevance and utility of the test need to be examined, as do the value implications of the test and the social consequences. These last three pieces of evidence are heavily reliant on the construct validity of the test. The test has to be shown to be reliably measuring the constructs the test is based on before the relevance and utility, value implications, and social consequences of the test could be discussed. Once all of the pieces of evidence have been gathered and investigated, justifications for testing outcomes can be demonstrated indicating that decisions made using test results are valid. Although the relevance and utility, value implications, and social consequences haven't been investigated, they can be discussed. A discussion of each of these follows with some suggestions for further investigation. Once further investigation has been done, these pieces of evidence could all be used to demonstrate the validity of the test for the purpose it was

designed—a test of ESL academic reading ability used to make placement decisions concerning which international students need further reading instruction.

Relevance and utility

To provide additional support for the evidence for the validity of the test, an investigation of the relevance and utility of this test as a test of ESL academic reading ability would need to be done. This test was designed because there is a need to separate those incoming international students who are ready to handle the reading they will encounter at ISU from those who are not. The students who aren't prepared to handle the reading receive further instruction which is designed to prepare them for future ESL reading. This is the context within which this test is used. The relevance and utility of a test use deals with the usefulness of the test and helps us to determine if a test is meeting its objectives within its context.

The relevance of the test used in this study is that it gives us an indication of a student's ESL academic reading ability. With this information, we can determine the best course of action for the student. This is done by comparing the student's test score with the cut-off score that has been set for the test. The cut-off score is a score that students must achieve to pass the test. Generally, to set a cut-off score, past experience is used to determine what level of mastery is required for students to pass a test. For this particular testing situation, the cut-off score was set according to how many positions were available in the ESL reading class (the cut-off score ended up being set at around the tenth percentile). Students who do not achieve the cut-off score are placed in the ESL reading class. This becomes the utility of the test—that of providing results from which decisions can made concerning the placement of international students into an ESL academic reading course. The objective of the

test, then, is to provide us with an indication of students' ESL academic reading abilities in order to separate those individuals who need further instruction in ESL academic reading skills from those who don't. The distribution of the scores allows this to be done. Students whose scores were below the cut-off score are identified and placed in the ESL reading class. Upon questioning the Placement Coordinator and the instructor who taught the ESL reading class, I found that they both agreed the test did a good job of separating out individuals who needed further instruction.

Value implications

In order to more accurately assess the justifications of the outcomes of a test, the degree of agreement between the underlying theorized constructs used for producing the test and the judgment of professionals concerning the hypothesized constructs needs to be established. The higher the agreement between professional opinion and hypothesized construct, the more value the interpretation of the test outcomes will have and the more valid will be the judgments made from the interpretations of the outcomes. In order to examine the value implications of the justifications of the outcomes of the test created for the study at hand, attention would be paid to the agreement between professionals opinions of reading comprehension and the theory underlying the construction of the test.

A high level of agreement between the hypothesized constructs underlying the creation of the test and the opinions of the professionals involved with making this test would result in a high level of value placed on the decisions made through the interpretations of the test results.

Social consequences

The social consequences of the test result from the use of the interpretations of test scores. The scores are used in judging whether or not a student is in need of further training in ESL academic reading. Failure of the student to achieve the cut-off score results in that student being placed in an ESL academic reading class. The social consequences are that those who need further instruction receive that instruction and those who equal or exceed the cut-off score are allowed to pursue their coursework without the additional time and cost of an ESL academic reading class. Those who really do need additional training receive it and will benefit by being better prepared to handle the academic readings they are about to encounter in their coursework.

An additional social consequence arises out of the fact that the test is designed to measure four different components of the reading process. Although the test is not designed as a diagnostic test, results from the test may give clues as to where a student is having difficulties. These clues, in turn, may be helpful in planning a program to help that student improve his/her academic reading skills.

Once the relevance and utility, value implications, and social consequences of the test have been investigated, they can be combined with the evidence used to investigate the construct validity in order to demonstrate the validity of test use.

With all of the pieces of evidence that are involved with the validation process, it is easy to see why the validity of the test for its purpose applies only to this one testing situation. Any of the pieces of evidence gathered could (and likely will) undergo changes over the course of time. As was mentioned earlier, different students take the test every semester, there are changes in the ESL committee from testing to testing which can influence expert opinions of what ESL academic reading

is, and continuing research may reveal the test isn't actually testing only ESL academic reading ability.

What has been learned

What has been learned from this study is that when designing a test, it is very important to have a clear definition of the test context and the test purpose as these will influence the construct for the test. It is also important to have a construct for the test that has support from those associated with the test development and implementation process. Without this, the validation process isn't possible.

It is also important to remember that the new view of validity does not consider a test to be valid for all time; instead, a test's validity must be demonstrated each and every time it is given. Although validity has been demonstrated for this test use, it cannot be assumed to be valid for each and every testing situation. Every testing situation will require the same process for demonstrating validity that has been used for the study at hand.

Postmodernism and language testing: Where do we go from here?

A concern that should be considered in the future of language testing is the effects that postmodernist theories such as such as social constructionism, deconstruction, philosophical hermeneutics, and externalism will have on the future of testing reading comprehension. Slowly but surely, the postmodernist movement is gaining strength in many of our major institutions. Postmodernist theories go against traditional positivist theories upon which many language tests are based. The result is possible clashes between postmodernist thought (with its breakdown between subjective and objective) and the traditional positivist need for objectivity.

Social constructionist theory is based in the notion that the knowledge an individual has is a result of the society in which that individual lives or participates (see Bruffee (1984) for a more detailed explanation of the social construction of knowledge). In essence, we are socially constructed and the society we come from influences how we interpret discourse. Deconstructionist theory places emphasis on examining not just what is in a text, but also what is not in a text. The theory hypothesizes that for every idea in a text, there is an opposing idea (or ideas) that exists but is not stated (Crowley, 1989). The text is a surface idea that has numerous deep ideas associated with it. Because of this, a true interpretation of a text is not possible since we cannot know for sure what ideas an author wants the audience to interpret. Philosophical hermeneutics states rather than knowing for sure how to interpret the meaning of a text, we make guesses (Crusius, 1991). We search for meaning that matches the meaning intended by the author. Unlike deconstruction with its notion that a text can never be properly interpreted, philosophical hermeneutics offers at least a chance that some truth for a text can be negotiated between an author and audience. Externalism goes somewhat beyond philosophical hermeneutics. The theory is a recent theory that is still in the development stage (I thank Thomas Kent (personal interview) for information pertaining to externalism). Externalism posits that all communication is a hermeneutical (interpretive) process; in essence, communication is a hermeneutical guessing game. We are constantly making guesses (and this is the best we can do) as to the real meaning behind an interlocutor's utterance.

Postmodernist theories are based heavily on the breakdown of subjectivity and objectivity. This is in strong contrast to traditional positivist views that discourse can be observed objectively. What these postmodernist theories mean to

reading comprehension testing is that we as creators of reading tests cannot be sure ourselves of how a passage is supposed to be interpreted. How can we be sure if we are testing the "true" interpretation of a text if we as the test creators are not guaranteed to be interpreting passages properly? How can we be sure that test subjects don't have a different but equally "truthful" interpretation of a passage—and should we penalize subjects for this?

Answers to these questions and others like them still aren't clear. There is much, much research to be done to answer questions such as these. Hopefully, developments such as those that have recently been seen in the validation process will help us to figure out answers to questions such as those above and allow us to create more accurate (and authentic) reading tests.

REFERENCES

Alexander, J. E. (1979). *Teaching Reading*. Boston: Little Brown and Company.

Alexander, J. E., & Heathington, B. S. (1988). *Assessing and Correcting Classroom Reading Problems*. Glenview, IL: Scott, Foresman and Company.

American Psychological Association (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Aronson, E., & Farr, R. (1988). Issues in Assessment. *Journal of Reading, 32* (2), 174-77.

Bachman, L. F. (1990). *Fundamentals Considerations in Language Testing* Oxford: Oxford Press.

Brown, J. D. (1990). Testing in Language Programs. Unpublished Manuscript, Department of ESL, University of Hawaii at Manoa.

Bruffee, K. A. (1984). Collaborative Learning and the "Conversation of Mankind". *College English, 46* (7), 635-652.

Carrell, P. L. (1984). The Effects of Rhetorical Organization on ESL Readers. *TESOL Quarterly, 18* (3), 441-69.

Chapelle, C. A. (1994). Are C-tests Valid Measures for L2 Vocabulary Research? *Second Language Research, 10* (2), 157-87.

Clark, M., & Silberstein, S. (1977). Toward a Realization of Psycholinguistic Principles for the ESL Reading Class. *Language Learning, 27*, 135-54.

Coady, J. (1979). A Psycholinguistic Model of the ESL Reader. In R. Mackay, B. Barkman, & R. R. Jordan (Eds.), *Reading in a Second Language* (pp. 5-12). Rowley, MA: Newbury House.

Connor, U. (1984). Recall of Text: differences Between First and Second Language Readers. *TESOL Quarterly, 18* (2), 239-56.

Crowley, S. (1989). *A Teacher's Introduction to Deconstruction* Urbana, IL: NCTE.

Crusius, T. W. (1991). *A Teacher's Introduction to Philosophical Hermeneutics* Urbana, IL: NCTE.

Davey, B., & Macready, G. B. (1985). Prerequisite Relations Among Inference Tasks for Good and Poor Readers. *Journal of Educational Psychology, 77* (5), 539-52.

Davis, F. B. (1944). Fundamental Factors of Comprehension in Reading. *Psychometrika, 9,* 185-97.

Davis, F. B. (1968). Research in Comprehension in Reading. *Reading Research Quarterly, 3,* 499-545.

Davis, F. B. (1972). Psychometric Research on Comprehension in Reading. *Reading Research Quarterly, 6,* 628-78.

Dechant, E. (1991). *Understanding and Teaching Reading: An Interactive Model.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Freedle, R., & Kostin, I. (1993). The Prediction of TOEFL Reading Item Difficulty: Implications for Construct Validity. *Language Testing, 10* (2), 133-70.

Goodman, K. (1967). Reading: A Psycholinguistic Guessing Game. *Journal of the Reading Specialist, 6,* 126-35.

Grabe, W. (1989). Literacy in a Second Language. *Annual Review of Applied Linguistics, 10,* 145-62.

Grabe, W. (1991). Current Developments in Second Language Reading Research. *TESOL Quarterly, 25* (3), 375-406.

Gray, W. S. (1917). Studies of Elementary School Reading Through Standardized Tests. *Supplementary Educational Monographs,* no. 1. Chicago, Ill: University of Chicago Press.

Howards, M. (1980). *Reading Diagnosis and Instruction: An Integrated Approach.* Reston VA: Reston Publishing Company.

Johns, J. L. (1986). *Handbook for Remediation of Reading Difficulties.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Johnson, M. S. (1946). Factors in Reading Comprehension. *Education Administration and Supervision, 35,* 385-406.

Kennedy, E. C. (1977). Classroom Approaches to Remedial Reading. (2nd ed.) Itasca, IL: F. E. Peacock Publishers, Inc.

Kent, T. (1995). Personal Interview.

Koda, K. (1988). Cognitive Process in Second Language Reading: Transfer of L1 Reading Skills and Strategies. *Second Language Research, 4* (2), 133-56.

Laufer, B. (1992). Reading in a Foreign Language: How Does L2 Lexical Knowledge Interact with the Reader's General Academic Ability? *Journal of Research in Reading, 15* (2), 95-103.

Lewis, C. M. (1987). Vocabulary, Sentences and Words: Testing for Agreement between Two Recent Measures of Reading Performance and Receptive Vocabulary. *Educational Psychology 7* (2), 129-32.

Lyons, K. (1984). Criterion Reference Reading Comprehension Tests: New Forms with Old Ghosts. *Journal of Reading, 27* (4), 293-98.

Messick, S. (1980). Test Validity and the Ethics of Assessment. *American Psychologist, 35* (11), 1012-27.

Messick, S. (1989). Validity. In R. L. Linn (Ed) *Educational Measurement* (3rd ed.). New York: Macmillan, 13-103.

Olsen, M. W. (1985). Text Type and Reader Ability: The Effects on Paraphrase and Text-Based Inference Questions. *Journal of Reading Behavior, 17* (3), 199-214.

Pettit, N. T., & Cockriel, I. W. (1974). A Factor Study on the Literal Reading Comprehension Test and the Inferential Reading Comprehension Test. *Journal of Reading Behavior, 6*, 63-75.

Readence, J. E., & Moore, D. W. (1983). Why Questions? A Historical Perspective on Standardized Reading Comprehension Tests. *Journal of Reading, 26* (4), 306-16.

Rost, D. H. (1993). Assessing Different Components of Reading Comprehension: Fact or Fiction. *Language Testing, 10* (1), 79-92.

Samuels, J., & Kamil, M. (1984). Models of the Reading Process. In P. D. Pearson, R. Barr, M. L. Kamil, & Mosenthal (Eds.), *The Handbook of Reading Research* (pp. 185-224). New York: Longman.

Singer, H., & et. al. (1988). English Classes as Preparation of Minimal Competency Tests in Reading. *Journal of Reading, 31* (6), 512-19.

Singleton, D., & Little, D. (1991). The Second Language Lexicon: Some Evidence from University-Level Learners of French and German. *Second Language Research, 7* (1), 61-81.

Smith, F. (1971). *Understanding Reading*. New York: Holt, Rinehart & Winston.

Spearitt, D. (1972). Identification of Subskills of Reading Comprehension by Maximum Likelihood Factor Analysis. *Reading Research Quarterly, 8*, 92-111.

Sternberg, R. J. (1991). Are We Reading Too Much Into Reading Comprehension Tests? *Journal of Reading, 34* (7), 540-545.

Thorndike, E. L. (1917a). Reading as Reasoning: A Study of Mistakes in Paragraph Reading. *Journal of Educational Psychology, 8,* 323-32.

Thorndike, E. L. (1917b). The Psychology of Thinking and the Case of Reading. *Psychological Review, 24,* 220-34.

Thorndike, E. L. (1917c). The Understanding of Sentences: A Study of Errors in Reading. The Elementary School Journal, 18, 98-114.

Thorndike, R. L. (1973-74). Reading as Reasoning. *Reading Research Quarterly, 11,* 185-88.

Thorndike, R. L., & Hagen, E. (1977). *Measurement and Evaluation in Psychology and Education* (4th ed.). New York: Wiley.

Thurstone, L. L. (1946). Note on a Reanalysis of Davis' Reading Tests. *Psychometrika, 11,* 185-88.

Vacca, R. T. (1980). A Study of Holistic and Subskill Instructional Approaches to Reading Comprehension. *Journal of Reading, 23,* 512-18.

Vernon, P, E. (1962). The Determinants of Reading Comprehension. *Educational and Psychological Measurement, 22,* 269-86.

ACKNOWLEDGMENTS

Many, many thanks to the following:

My POS committee:
Carol Chapelle, Barb Schwarte, and Fred Duffelmeyer for being so patient with me, and **Dan Douglas** for filling in at the last minute.

Members of the ESL **Committee**:
for your help with creating the test.

The many **101C and 101D** students who volunteered to take the pilot tests.

The **Test Evaluation Center**:
for the many, many item analyses you ran for me.

The **English Department**:
for allowing me the opportunity to advance my education.

My lord and savior, **Jesus Christ**:
for seeing me through this project and helping me through the many rough moments.

APPENDIX A: GUIDELINES FOR GATHERING EPT READING PASSAGES

The purpose of this exam is to test for academic reading skills. The test will ask examinees to read several passages and answer questions related to the passages. Since we are interested in testing academic reading ability, we will want a sample of passages that is representative of academic reading materials. More specifically, we want materials that represent the types of materials students are likely to encounter here at ISU. The goal is to gather a minimum of 12 total passages. There are some basic guidelines we should follow when searching for appropriate reading materials.

- Avoid topics that may be considered offensive in other cultures. Although issues such as sex, religion, race, and gender are discussed frequently here in the US, these subjects may be considered taboo in other cultures.
- Avoid bias. We want to gather materials from as many different topics as possible to insure that no one has a distinct advantage due to background knowledge.
- Avoid topics that are common knowledge. This could allow for a set of questions that could be answered without actually reading the passages.
- Avoid topics that are too specific or too technical. This could result in a passage filled with topic specific or highly technical terminology making the passage too difficult or time-consuming to read.
- Look for passages that are from familiar topics, but discuss specific aspects of the topics.
- Look for passages with a clear main topic. Although it does not have to be explicitly stated, it should be clear from reading the passage what the main topic is.
- Passages should be limited to 250–400 words. Although the passages may be more than one paragraph long, there should be only one main topic.
- Use reading materials here at ISU. These materials may include textbooks and journals.
- Assume that since the examinees are entering ISU they are at least as intelligent as incoming freshmen. In other words, don't assume that lack of English skills means lack of intelligence. Search for passages that an average ISU students is likely to encounter.
- Reading passages for the test will generally be parts of larger reading passages. Be care to make sure that test passages contain complete ideas that do not draw upon material that isn't included in the part of the reading passage chosen.

Sample Passages

To further assist you in your search for appropriate passages, here are some sample passages that demonstrate an appropriate and an inappropriate passage. The first is

an example of the kind of passage to look for. The second is an example of a passage to avoid.

First Passage

The Wilson Phalarope

Like all sandpipers, the Wilson Phalaropes have long legs and beaks for walking and picking through shallow water on the hunt for food. Although not seabirds, they are highly specialized for aquatic life. While others of their family—sandpipers, for example—stay along the shore, phalaropes swim well on are at home on land and water.

The life cycle of the bird, named after the eminent 19th century naturalist Alexander Wilson, begins in May on its major breeding grounds in the prairie pothole country of the northern Great Plains of the United Sates and southern Canada. But the bird winters mainly on the high altitude lakes of the central Andes of South America.

After the female phalaropes lay the eggs, the males take all responsibility for incubating them and raising the chicks. Thus free of parental duties, the females start their southward migration early; by mid-June they are assembled in small flocks in preparation for the first leg of their journey. Taking wing, many head for Mono Lake, where they will prepare for their long nonstop flight to South America.

This passage, though somewhat short, is a good because there is a clear main topic. Although not many people are familiar with the specific topic of Wilson phalaropes, everyone is familiar with the more general topic of birds. Notice also that the vocabulary consists primarily of commonly used words with a few less commonly used words such as "eminent" and "aquatic" which can be used for vocabulary questions. Also, issues such as race and religion are not covered so the chances of offending the reader with the subject matter in this passage is quite low.

Second Passage

Mathematics

Real numbers are represented by symbols such as

25, 0, -3, .5, -.125, .333..., π

The set of counting numbers, or natural numbers, is the set {1, 2, 3, 4, ...}. The set of integers is the set {..., -3, -2, -1, 0, 1, 2, 3, ...}. A rational number is a number that can be expressed as a quotient a/b of two integers, where the integer b cannot be zero. Examples of rational numbers are 3/4, 5/2, and -2/3. Since $a/1 = a$ for any integer a, every integer is also a rational number. Real numbers that are not rational are called irrational. An example of an irrational number is π which equals the constant ratio of circumference to diameter of a circle.

Real numbers can be represented as decimals. Rational real numbers have decimal representations that either terminate or are non-terminating with repeating blocks of digits. For example, 3/4 = .75, which terminates; and 1/3 = .333..., in which the digit 3 repeats indefinitely. Irrational real numbers have decimal representations that neither repeat nor terminate. For example, $\pi = 3.14159...$. In practice, irrational numbers are generally represented by approximations.

Often, letters are used to represent numbers. If the letter used is to represent any number from a given set of numbers, it is referred to as a variable. A constant is either a fixed number, such as 5, or a letter that represents a fixed number.

In working with expressions or formulas involving variables, the variables may only be allowed to take on values from a certain set of numbers, called the domain of the variable. For example, in the expression $1/x$, the variable x cannot take on the value 0 since division by 0 is not allowed.

The second is the kind of passage to avoid. Even though most students have had at least some exposure to mathematics, this is an inappropriate passage because of the large number of technical terms. Also, questions concerning the content could be easily answered by a person familiar with math.

Guidelines for Test Items

There are three specific outcomes we want to examine: did the examinee comprehend the passage, can s/he make inferences from the material, and can s/he determine the meaning of specific words according to their use in the passage? In order to do this, we will need to ask a variety of test items.

- Comprehension Questions. These questions will ask the examinees to complete tasks such as identifying the main idea or recognizing and understanding some of the supporting ideas.
- Inference Questions. These questions will ask the examinee to use the material in the passages to answer questions regarding ideas, which would be logical inferences on the basis of what is stated, but not explicitly stated.
- Vocabulary Questions. These questions will ask the examinee to determine the meaning of the word according to the context in which it is used.

All test items will be in a multiple choice format with four possible answers. One of the possibilities will be the correct answer. The other three, which should be clearly wrong, will serve as distracters.

For each passage, we will want to create a total of three to five comprehension questions, three to five inference questions, and three to five vocabulary questions.

To further assist you in creating test items, samples of test items for the Wilson Phalarope passage are provided.

Comprehension Test Items

Select the statement which best expresses the main idea of this passage.

a. The Wilson phalarope is similar in many ways to the sandpiper.

b. By mid-June, the female phalaropes are ready to begin their migration to the Andes.
c. The Wilson phalarope is named after the eminent 19th century naturalist, Alexander Wilson.
d. The Wilson phalarope has a life cycle that begins in North America and continues in the Andes in the winter.

By mid-June, the female phalaropes have gathered for their journey. Their first destination is
a. Mono Lake.
b. southern Canada.
c. the central Andes of South America.
d. the northern Great Plains of the United States.

Inference Test Items

After reading this passage, we can conclude which of the following to be false:
a. The females have little contact with the chicks.
b. The females show a strong instinct for mothering the chicks.
c. The males migrate to South America later than the females do.
d. The males carry the responsibility for feeding and tending the chicks.

The diet of the Wilson phalarope would most likely include
a. berries or fruits.
b. seeds and grains.
c. small fish or seaweed.
d. small animals such as rats and mice.

Vocabulary Test Items

The best definition for the word "aquatic" in line 2 is:
a. things that eat fish.
b. things that live on land.
c. things that drink water.
d. things that live in or near water.
(The word "aquatic" might favor many Spanish speakers. This is something that should be considered when creating vocabulary questions.)

The word "winters" in line 7 is used in this essay to mean
a. migrates.
b. lives only in high altitudes.
c. lives more than one winter.
d. stays during the winter season.

APPENDIX B: SAMPLE PASSAGE

Cyclamen in all Their Infinite Variety

Variation within species is a good thing. It is the engine that drives evolution and is a constant reminder to us that Nature knows neither divisions nor categories: she is not a taxonomist. Some species show variation more than others. You can look all day for a significantly different daisy on your lawn, but examine a cyclamen and the chances are that it will be an individual with at least one characteristic that is different from all the others. It may be in the flower or the foliage, but it is there to be found.

So great are the horticulturally significant differences within *Cyclamen persicum* that they have given rise to the entire range of florists' cyclamen and have even allowed breeders to extract from its genes colors that do not occur in the wild. Pink, rose, lilac, purple and white occur in nature, but red does not. Similarly, no wild specimen attains the size of the larger florists' forms.

This tender species, which cannot be persuaded to live for long in the open, even in the mildest areas, is increasingly represented by florists' forms that are smaller and bear more resemblance to the exquisitely modeled plants found in its home area around the eastern Mediterranean and beyond. Its wonderful scent, lost almost entirely in breeding for spectacle, has returned in some of these more dainty versions.

The species itself, whose foliage is almost infinitely varied in its marbling, marking and shape, is all grace and elegance. The flower stalks are usually 10-15cm (4-6in) high, but may occasionally be as much as 20cm (8in). The petals of almost all cyclamen are reflexed, but in this species they are swept back with panache, each one long, slender and with a balletic, upward half twist. The flowers are fragrant.

1. The main idea of this passage is

   a. nature is not a taxonomist.
   b. there is much variation in cyclamen.
   c. some species have more variety than others.
   d. florists' cyclamen have a greater variety of colors.

2. Which of the following colors will one probably not find in cyclamen found growing around the eastern Mediterranean?

   a. Red.
   b. Rose.

c. Lilac.
d. White.

3. If someone were describing a flower to you, for which of the following characteristics would you say, "That's not a cyclamen!"?

a. Delicate.
b. Graceful.
c. Odorless.
d. Hardy.

4. "It" in the last line of the first paragraph (line 7) refers to

a. a daisy.
b. a cyclamen.
c. an individual.
d. a characteristic.

5. "Spectacle" in line 17 means

a. show.
b. science.
c. enjoyment.
d. experimentation.

6. Cyclamen bred artificially often differ from cyclamen in nature in that they are

a. larger.
b. marbled.
c. more dainty.
d. more fragrant.

7. In paragraph two, line 9, "have given rise to" means

a. have been grown for.
b. have been the basis for.
c. have increased the number of.
d. have made known the fact that.

8. "Breeders" in line 10 means people who

a. sell flowers.
b. like flowers.
c. raise flowers.
d. photograph flowers.